Supplemental Methods

Attention and Saliency Maps

In addition to the model's performance in predicting non-POAG and POAG classes, it is essential to evaluate the inference process and explain how the model reaches the outcome for each input case. For that purpose, visual explanations such as saliency and attention maps are very popular approaches incomputer vision applications.

Saliency maps (such as ScoreCAM, GradCAM++, and GradCAM rows in Figure 3) are generally achieved in a post hoc process via guided back-propagation and class activation maps. Saliency maps highlight those parts of the input that the AI algorithm is susceptible to. On the other hand, attention maps are learned by attention-based AI algorithms (such as DeiT) in the training process and reflect the algorithm's holistic knowledge of what is important in each input image. The transformer layers in DeiT help the model build an attention mechanism to identify the most important areas of the image for classification. This attention mechanism benefits from the spatial encoding of the small pixel patches in each image to identify the important regions.It is worth noting that although DeiT learns to calibrate its attention over different parts of the input image, this attention mechanism is not necessarily intuitive and interpretable to human experts.

To visualize the attention maps for DeiT (Transformer row in Figure 3), we looked closely at the transformer attention mechanism within the model. The model first transforms the input image into a 16×16 set of image patches. The model generates an attention tensor to define howthese patches should interact with each other to produce the final output of the model. The transformer contains 12 layers, and each layer includes 12 attention heads. The transformer directs the attention to the image patches throughout these layers. The attention map in the transformer is derived from its last layer, where we average the attention weights among all 12 heads. The attention map is then generated by normalizing the average of attention weights and highlighting the patches that are attended above a threshold.